(19) Europäisches Patentamt
Europ an Patent Office
Offic uropéen d br v ts

(11) EP 0 964 344 A2

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
15.12.1999 Bulletin 1999/50

(51) Int Cl.[6]: **G06F 17/30**

(21) Application number: **99304218.3**

(22) Date of filing: **28.05.1999**

(84) Designated Contracting States:
AT BE CH CY DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE
Designated Extension States:
AL LT LV MK RO SI

(30) Priority: **02.06.1998 GB 9811744**

(71) Applicant: **SHARP KABUSHIKI KAISHA**
Osaka 545 (JP)

(72) Inventors:
- **Ijdens, Jan-Jaap**
  Gloucester Green, Oxfordshire OX1 2DF (GB)
- **Poznanski, Victor**
  Sandford on Thames, Oxford (GB)

(74) Representative: **Robinson, John Stuart**
Marks & Clerk,
4220 Nash Court,
Oxford Business Park South
Oxford OX4 2RU (GB)

(54) Method of and apparatus for forming an index, use of an index and a storage medium

(57) In order to form an index for a plurality of documents, each of the words or collocations are identified (24,25) and applied to a multilingual resource (11) or a thesaurus (12). Linguistically related but different terms, such as translation into a different language or synonyms from a thesaurus, are generated (28). These equivalent terms are ordered (32) in accordance with the likelihood of being correct or appropriate based on information contained in the multilingual resource (11) or thesaurus (12). Each of the words or collocations in the documents and each of the most likely equivalent terms is then formed into an indexing feature by appending an identifier for the or each document in which it occurs (34). This provides an index which may be used to locate documents using terms contained in the document but also using derived terms such as translations into a different natural language and synonyms.

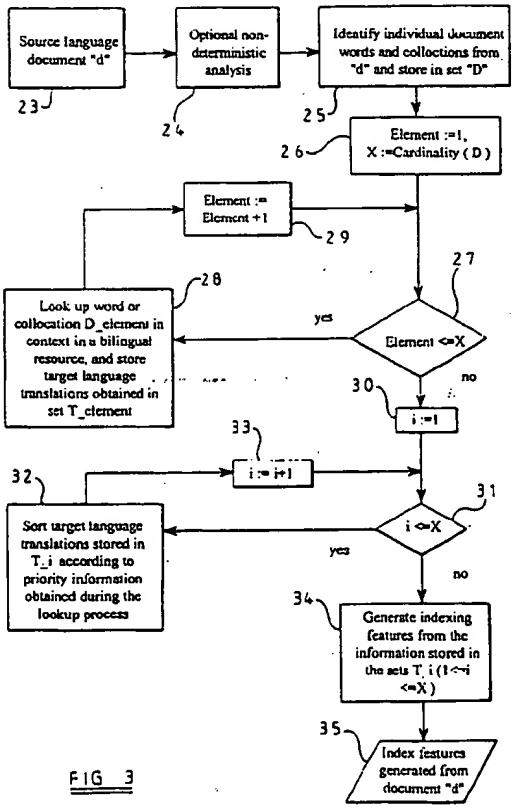

FIG 3

EP 0 964 344 A2

Description

[0001] The present invention relates to a method of and an apparatus for forming an index. Th invention also relates to a storage medium storing a program for performing the method, an index, a storage medium containing th index and the use of the index to access docum nts.

[0002] The techniques disclosed herein may be used for information management. Examples of such applications include information retrieval systems, such as search engines, for accessing information on the internet or in office information systems, information filtering applications (also known as information routing systems) and information extraction applications.

[0003] There are many data bases which contain documents in machine-readable form and which can be accessed to locate and retrieve information. Similarly, there are various known techniques for locating documents on the basis of subject matter. One example of this is the collection of published patent specifications. All patent specifications are indexed according to subject matter when the specification is published in accordance with the International Classification. The content of each patent specification is analysed in accordance with the International Classification and the relevant classification numbers for the subject matter form part of the heading of both the printed patent specification and the machine-readable form.

[0004] In order to locate patent specifications, or indeed other documents, whose collections are similarly classified according to subject matter, it is necessary to select the correct international class and to apply this to a searching system. The searching system then locates all patent specifications which have been classified in the same class. However, a disadvantage of this system is that efficient use requires familiarity with and experience of using the International Classification system. Also, this technique relies on correct classification of patent specifications. Inexperienced use can result in relevant patent specifications being missed whereas incorrect classification can prevent a relevant patent specification from ever being located by this technique.

[0005] Another known technique for information retrieval relies on the selection of keywords which are then used to search for relevant documents such as patent specifications. In this case, it is necessary to identify words which are likely to appear in the relevant documents but which are unlikely to appear in irrelevant documents. Searching using keywords then reveals all documents which contain the keywords or combination of keywords.

[0006] There are several difficulties with this technique. For instance, in the case of subject matter without well-defined or standard terminology, it may be difficult or impossible to select all keywords which might identify relevant documents. On the other hand, the use of more general keywords can lead to the disclosure of very large numbers of documents many of which are irrelevant. Further, such keywords can only be used for documents which are in the same language or which have been completely or partially translated or abstracted into the language of the keywords. The effectiveness of this technique in locating documents in other languages may therefore be poor or non-existent.

[0007] D.A. Hull and G. Greffenstette, "Querying across Languages: a Dictionary-Based Approach to Multilingual Information Retrieval", 19th Annual International Conference on Research and Development in Information Retrieval (SIGIR'96), pages 49-57, 1996 and D.W. Oard and B.J. Dorr, "A Survey of Multilingual Text Retrieval", Technical Report UMIACS-TR-96-19, University of Maryland, Institute for Advanced Computer Studies, April 1996, disclose two techniques for performing multilingual information retrieval, one based on document translation and the other based on query translation. In each case, each translation is to be performed by a machine translation system. Thus, in the cas of document translation, a machine translation system is used to translate all of a collection of documents into a target language so that queries for locating and retrieving information, for instance based on the keyword technique described hereinbefore, may be performed in the source (document) language or in the target language. In the other technique, the documents are not translated but each query is translated into the source or document language and the translations are used to search the document collection.

[0008] A disadvantage with query translation is that queries often comprise a few words and may not even be in a sentence context. Thus, automatic linguistic processing of such queries can be difficult and may lead to unsatisfactory results, such as failure to locate relevant documents and location of irrelevant documents.

[0009] The use of automatic machine translation to translate whole collections of documents to form an index is also problematic. The resources required in terms of computing time and additional storage medium capacity make this technique unattractive. Although such processing need not be performed in real time and, in particular, is not required as part of each information retrieval request, substantial resources are necessary and there may be a continuing requirement as further documents are added to the collection. Translation into several target languages multiplies the resource requirements.

[0010] Machine translation systems also perform tasks which are not useful to information retrieval and, in particular, to the forming of a multilingual index. For instance, in addition to translating words and groups of words contained in documents, machine translation systems also attempt to produce a good quality translation which is readable for human beings. If the translation is merely required for indexing, functions such as correct word ordering in the target language are unnecessary and are therefore wasteful of computing resources.

[0011] A further disadvantage with machine translation systems when used to translate documents into a target language for indexing purposes is that the effectiveness of the index may be seriously compromised. Some machine translation systems generate a single preferred translation of an input text. In other words, such systems attempt to identify and produce a single translation which is judged according to automatic criteria within the system as the best translation. If that translation is incorrect, then retrieval of information based on the incorrect translation will be ineffective because relevant documents may not be located and irrelevant documents may be located.

[0012] Other machine translation systems attempt to generate all possible translations of input text. Thus, even if the correct translation is present, there may be many other translations which are inappropriate or wrong. The use of such translations for information retrieval results in the generation of spurious matches on queries posed to the system so that very large numbers of irrelevant documents may be located together with the relevant documents.

[0013] WO 97/08604 discloses a document retrieval system in which documents and queries are converted or reduced to a common language-independent conceptual representation.

[0014] EP 0 813 160 discloses a technique for accessing a database of a type in which each entry comprises a main or head word with one or more associated words as subsidiary entries. If a word of a query matches the head word of an entry, the words of the query are checked for the presence of words in the subsidiary entries.

[0015] EP 0 304 191 discloses a searching system in which the words of a query are processed to find equivalent words for use during searching.

[0016] According to a first aspect of the invention, there is provided a method of forming, for a plurality of documents, an index comprising indexing features, the method comprising the steps of:

identifying each of at least some of the terms present in the documents;

generating from each identified term at least one equivalent term which is different from but linguistically related to the identified term;

forming for each of the identified terms a first indexing feature comprising the identified term and an identifier of the or each document in which the identified term occurs;

forming for each of the equivalent terms a second indexing feature comprising the equivalent term and an identifier of the or each document in which the identified term to which the equivalent term is equivalent occurs; and

forming an index comprising the first and second indexing features.

[0017] The expression "term" used herein means an individual word, a group of linked words which occur adjacent each other in a document (continuous collocation), or a group of words which are linked to each other but which are divided into at least two subgroups of words separated in a document by one or more words which are not members of the group (non-continuous collocations).

[0018] The expression "identifier" as used herein is any means for identifying one or more locations of a term, for instance a heading or arbitrary serial number of a document containing the term. The expression "indexing feature" as used herein means a term and an identifier.

[0019] The expression "linguistically related" as used herein means a term which has the same, a similar or related meaning. For instance, linguistically related terms include synonyms, more general terms and more specific terms in the same (natural) language and translations into a different (natural) language.

[0020] Although the documents may be in any type of language, such as a high level computer programming language, the documents are preferably natural language documents.

[0021] The generating step may comprise accessing a thesaurus with each identified term and the equivalent terms may be synonyms of the identified terms, more general terms than the identified terms and more specific terms than the identified terms.

[0022] The generating step may comprise accessing a multilingual resource with each identified term and the equivalent terms may be translations of the identified terms, more general terms than the identified terms and more specific terms than the identified terms.

[0023] The multilingual resource may comprise a glosser. The glosser may be a limited non-deterministic glosser. The glosser may form a plurality of translations of at least one of the identified terms and may assign to each translation a priority according to the likelihood of the translation being correct.

[0024] The multilingual resource may comprise a bilingual dictionary.

[0025] The multilingual resource may comprise a machine translation system.

[0026] The identifying step may be performed by a part of speech tagger.

[0027] According to a second aspect of the invention, there is provided an apparatus for forming, for a plurality of

documents, an index comprising indexing features, characterised by comprising:

means for identifying each of at least some of the terms present in a document;

means for generating from each identified term at least one equival nt term which is differ nt from but linguistically related to the identified term;

means for forming for each of the identified terms a first indexing feature comprising the identified term and an identifier of the or each document in which the identified term occurs;

means for forming for each of the equivalent terms a second indexing feature comprising the equivalent term and an identifier of the or each document in which the identified to which the equivalent term is equivalent occurs; and

means for forming an index comprising the first and second indexing features.

[0028] The generating means may comprise a thesaurus and the equivalent terms may be synonyms of the identified terms, more general terms than the identified terms and more specific terms than the identified terms.

[0029] The identifying means and the generating means may comprise a multilingual resource.

[0030] The multilingual resource may comprise a glosser. The glosser may be a limited non-deterministic glosser. The glosser may be arranged to form a plurality of translations of at least one of the identified term and to assign to each translation a priority according to the likelihood of the translation being correct.

[0031] The multilingual resource may comprise a machine translation system.

[0032] The generating means may comprise a bilingual dictionary.

[0033] The identifying means may comprise a part of speech tagger.

[0034] The apparatus may comprise a programmed data processor.

[0035] According to a third aspect of the invention, there is provided a storage medium characterised by containing a program for controlling a data processor to perform a method according to the first aspect of the invention.

[0036] According to a fourth aspect of the invention, there is provided an index characterised by being formed by a method according to the first aspect of the invention or by an apparatus according to the second aspect of the invention.

[0037] According to a fifth aspect of the invention, there is provided a storage medium characterised by containing an index according to the fourth aspect of the invention.

[0038] According to a sixth aspect of the invention, there is provided use of an index according to the fourth aspect of the invention to access the documents.

[0039] It is thus possible to form an index to a collection of documents having indexing features which are not restricted to the terms which occur in the documents. By making use of the thesaurus entries, synonymic, more general and more specific terms may be added to the index to increase the likelihood that an arbitrary query will locate a relevant document during information retrieval. By using multilingual resources, indexing may be performed in an efficient and effective manner in languages other than the source or document language.

[0040] Although any type of multilingual resource may be used, light-weight cross-linguistic glossing systems have advantages. Such a glossing system uses limited non-determinism to generate plausible target language translations to be used in indexing features. Such glossing systems or glossers may be of the type disclosed in EP 0 813 160 and GB 2 314 183, the contents of which are incorporated herein by reference. This type of glosser is capable of identifying and translating sequential (continuous) and non-sequential (non-continuous) collocations which are indexed by a headword. Further, this system can be used to ascribe priorities to alternative translations in such a way that consistent translations of complete sections of text are always obtained irrespective of which of several translations of a word or collocation is in fact selected. Further, the prioritising of alternative translations allows a limited number of such translations to be used, for instance based on the priority information.

[0041] Such glossers are more efficient than machine translation systems. An index merely requires the identification and translation of terms and does not require other processing steps such as parsing and generation of a readable translation as provided by machine translation systems. Thus, the use of glossing is computationally more efficient in that substantially less computational time is required.

[0042] The use of a glosser can overcome the problems associated with selection by a machine translation system of a single most likely, but perhaps incorrect, translation and the selection of all possible translations including those which are incorrect and may b ntirely inappropriate for indexing purposes. By using non-deterministic techniques, a limited number of most likely translations of the t rms can be provided. There is a very high probability that this limited number of translations selected from all possible translations will include th best or correct translation. Accordingly, accessing documents using indexes formed in this way provides a high probability of locating all relevant documents while reducing the numb rs of irrelevant docum nts which might otherwise be located.

[0043] The invention will be further described by way of example, with reference to the accompanying drawings, in which:

Figure 1 is a block schematic diagram of an apparatus for forming an index constituting an embodim nt of the invention; and

Figures 2 and 3 are flow diagrams illustrating a method of forming an index constituting an embodiment of th invention and performed by the apparatus shown in Figure 1.

[0044] Figure 1 shows an apparatus for forming an index to a plurality of documents in machine-readable form stored in a document store 1, such as a magnetic disc or an optical storage medium such as a CD-ROM. The apparatus is of the programmed data processor type, such as a computer, and comprises a programmable data processor 2 provided with an input interface 3, such as a keyboard and mouse, and an output interface 4, such as a display and printer. Th data processor 2 has a "working memory" in the form of random access memory (RAM) 5 for temporarily storing data during data processing. A non-volatile read/write memory 6 is provided for storing data which are required to be retained, for instance, when the power supply to the apparatus is switched off A program memory 7 in the form of a read-only memory (ROM) contains a program for controlling operation of the data processor 2.

[0045] The apparatus may also be provided with other memory devices. For instance, these may comprise suitabl drives for CD-ROMs 8, floppy discs 9, and digital video discs (DVDs) 10. These devices may be of the read-only type or, for instance in the case of floppy discs 9, the read/write type. Such devices may provide the document store 1 and may provide an output medium for the apparatus. For instance, the index formed by the apparatus may be written to any of the storage media 8,9,10 shown in Figure 1.

[0046] The program memory 7 contains the aforementioned program which is executed by the data processor 2 and/or the multilingual resource 11 and/or the thesaurus 12 in order to carry out the various operations described herein. The program may be written in any of a variety of known computer languages as will be readily apparent to those having ordinary skill in the art of computer programming. Hence, further detail regarding the specific code itself has been omitted for the sake of brevity.

[0047] A multilingual resource 11 and a machine-readable thesaurus 12 are shown as individual devices in Figure 1. However, these devices may be embodied within the components of the apparatus already described. For instance, any of the memories and devices 7 to 10 may contain the data and the memory 7 may contain programs for performing the operations of the multilingual resource and/or the thesaurus 12.

[0048] The multilingual resource 11 shows four resources which may be used during operation of the apparatus. A document glosser 13 is a "device" which labels an "ordered" plurality of source language words or collocations (groups of words) with target language translations. The glosser is preferably of the type which orders the translations of each word or collocation in order of likelihood of being the "correct" translation. This is preferably of the limited non-deterministic type, for instance as disclosed in EP 0 813 160 and GB 2 314 183.

[0049] Although the document glosser 13 is the preferred type of multilingual resource for the apparatus, other types of resource are illustrated in Figure 1. Thus, the multilingual resource 11 may comprise a machine translation system 14. A suitable machine translation system is disclosed in W. John Hutchins and Harold L. Somers, "An Introduction to Machine Translation", Academic Press, 1992, ISBN 0-12-362830-X, the contents of which are incorporated herein by reference. A machine translation system performs deeper source language analysis than a glosser and also performs steps normally referred to as "generation", which attempts to put the target language translation of source language words or collocations into the correct grammatical order for the target language and to generate the correct inflections, etc. As described hereinbefore, such machine translation systems perform more processing and require more resources than a glosser but may be used as the multilingual resource 11 in appropriate circumstances.

[0050] The multilingual resource 11 may comprise a bilingual dictionary 15 of the machine-readable type. For instance, the source language text may not be processed but may simply be divided into words, and possibly collocations, which are then used to access the dictionary 15 to provide word-by-word translations of the text.

[0051] The multilingual resource 11 may comprise a part of speech tagger 16. Such a "device" performs limited grammatical analysis of the source language text to determine the part of speech of each word. The result of this limited analysis may then be applied to the bilingual dictionary 15 so as to provide an improved word-by-word translation than may be obtained by using the dictionary 15 alone.

[0052] Figures 2 and 3 illustrate a method of forming an index which may be performed by the apparatus shown in Figure 1. Th document store 1 contains a collection S of documents in a source language such as a natural language. For the purpos of illustration only, operation will be describ d for the case where the documents of the collection S are in English and an English and Dutch index is required. However, the method may b used to provide a single-language index making us of the thesaurus 12 so as to provide indexing features based not only on words and collocations contained in the documents but also based on words and collocations in the same language derived from

accessing the thesaurus 12 but not necessarily present in th documents.

[0053] In a step 20, a parameter "d" is set to a value of one and a parameter N is set equal to the cardinality of the document collection S ie. to the number of documents in the collection. A step 21 tests whether "d" is less than or equal to N and, if so, performs a routine 22, which applies a cross-linguistic indexing feature generator on a document identified as "d" and which is shown in mor detail in Figur 3.

[0054] The source language document d is shown at 23 and is supplied to an "optional" non-deterministic analysis at a step 24 and then to a step 25, which identifies individual document words and collocations from the document d and stores them in a set D. The step 24 is performed in turn on each sentence of the document d and represents a non-deterministic analysis of the source language of the sentence. The analysed sentence is passed to the step 25, which identifies individual words or collocations which potentially have translation equivalents in the target language. The steps 24 and 25 are performed by the document glosser 13.

[0055] In a step 26, a parameter "element" is set to a value of one and a parameter X is set to the value of the cardinality of the set D ie. the number of words and collocations in the set D. A step 27 tests whether "element" is less than or equal to X and, if so, a step 28 is performed. In the step 28, the word or collocation identified as $D_{element}$ is looked up in a bilingual resource, such as the dictionary 15. Each of the possible translations obtained from the dictionary is stored in a set $T_{element}$. The context of the word or collocation is taken into account so as to ensure that the translations into the target language make sense. For instance, this takes account of the possibility that certain decisions made about the translation of one part of a sentence may affect translations of other parts of the sentence.

[0056] In a step 29, the parameter element is incremented by one and the step 27 is performed again. This loop continues until all of the words and collocations in the set D have been translated, after which a step 30 is performed.

[0057] The step 30 sets a parameter i to a value of 1 and a step 31 tests whether i is less than or equal to X. If so, a step 32 sorts the target language translations stored in the set Ti according to priority information obtained during the step 28 from the bilingual dictionary. Thus, the step 32 provides a prioritising or ordering of each set of translations corresponding to a source language word or collocation. A technique for deriving such priority information is disclosed in EP 0 813 160 and GB 2 314 183.

[0058] A step 33 increments i by one and the step 31 is performed again. The loop continues until all target language translations have been sorted by the step 32, after which a step 34 generates indexing features from the information stored in the sets Ti for $1 \leq i \leq X$. In particular, the step 34 selects the most likely translations using the ordering generated in the step 32. The source words and collocations and the remaining target language translations are then arranged as indexing features by appending an identifier of the document d in which they were contained or from which they were derived. The resulting indexing features are shown diagrammatically at 35.

[0059] As shown in Figure 2, the target language features are added to a target language index T in a step 36. The parameter d is incremented by one in a step 37 and the step 21 is performed again. This procedure is repeated until all of the source language documents have been processed, at which point the target language index T is returned as shown at 38 together with the source language index to the output interface 4 and/or to any of the storage media 8, 9, 10.

[0060] A specific example to illustrate this method will now be described. In this specific example, the source language documents are in English and it is required to be able to access the documents in English or Dutch. The documents are therefore applied one at a time by the steps 20, 21 and 37 shown in Figure 2 to the analysis shown in Figure 3. For instance, the operation of the analysis illustrated in Figure 3 will be described with reference to a document having an identifier number #8. Document #8 comprises English sentences which are analysed one at a time. As an example, the following English sentence occurs in the document: "air passes out of the furnace".

[0061] The analysis step 24 identifies that "air" could be a noun or verb, "passes" could be a plural noun or the third person of a verb etc. The step 25 identifies all words and collocations in the sentence to provide the following analysis.

```
air_NOUN
air_VERB
pass_VERB
pass_NOUN
pass_VERB out_PREP
out_PREP
out_PREP of_PREP
of_PREP
the_DET
furnace_NOUN
```

[0062] The step 28 looks up the words and collocations in the bilingual dictionary or lexicon to derive Dutch translations as follows ("<none>" means that it is possible to give the word or collocation no translation):

air_NOUN	→ {lucht, hemel}
air_VERB	→ {luchten, uiten}
pass_VERB	→ {doorgeven, halen}
pass_VERB	→ {pas, kaart, voldoende}
pass_VERB out_PREP	→ {doorvoeren, flauwvallen}
out_PREP	→ {uit, buiten, extern van}
out_PREP of_PREP	→ {uit, buiten}
of_PREP	→ {<none>, van}
the_DET	→ {de, het}
furnace_NOUN	→ {oven, fornuis}

[0063] The step 32 orders the target language translations in order of likelihood of being correct and assigns them to the input sentence as follows:

air	→ [lucht, luchten, hemel, uiten]
pass	→ [doorvoeren, doorgeven, pas, kaart, voldoende, halen, flauwvallen]
out	→ [uit, buiten, extern van]
of	→ [<none>, van]
the	→ [de, het]
furnace	→ [oven, fornuis]

[0064] The step 34 generates the indexing features by applying the limited non-determinism ie. selecting the most likely translations, and associating the source language words and collocations and the target language translations with the identifier (#8) of the document currently being analysed as follows:

("air",#8)
("lucht",#8)
("luchten", #8)
("doorvoeren",#8)
("pass",#8)
("doorgeven",#8)
("pas",#8)
("kaart",#8)
("out of',#8)
("uit",#8)
("the",#8)
("de",#8)
("het",#8)
("furnace",#8)
("oven",#8)

[0065] Once all of the documents have been analysed in this way, the final index is provided, for example in a storage medium, and is of the following form:

["aardvark"]	→ #1,#17,#21,#47,#109
["air"]	→ #5,#8,#87
...	... ...
["out of']	→ #8,#10
...	... ...
["doorvoeren"]	→ #1,#8,#79
...	... ...
["zebra"]	→ #9,#10,#94,#187

[0066] Accordingly, when it is desired to retrieve information from the document collection, queries in either English or Dutch may be applied to the index by means of an information retrieval system. These queries may be in th form of words or collocations related to the subject matter to be searched. The information retrieval system applies th s to the index and, if match s with the indexing features are found, the relevant document number or numbers are returned so as to identify th document or documents which are likely to contain the subject matter of inter st.

Claims

1. A method of forming, for a plurality of documents (1, 23), an index comprising indexing features, the method comprising the steps of:

   identifying (24, 25) each of at least some of the terms present in the documents (1, 23);

   generating (28) from each identified term at least one equivalent term which is different from but linguistically related to the identified term;

   forming (34) for each of the identified terms a first indexing feature comprising the identified term and an identifier of the or each document in which the identified term occurs;

   forming (34) for each of the equivalent terms a second indexing feature comprising the equivalent term and an identifier of the or each document in which the identifier term to which the equivalent term is equivalent occurs; and

   forming an index comprising the first and second indexing features.

2. A method as claimed in claim 1, characterised in that the documents (1, 23) are natural language documents.

3. A method as claimed in claim 1 or 2, characterised in that the generating step (28) comprises accessing a thesaurus (12) with each identified term and the equivalent terms are synonyms of the identified terms, more general terms than the identified terms and more specific terms than the identified terms.

4. A method as claimed in claim I or 2, characterised in that the generating step (28) comprises accessing a multilingual resource (11) with each identified term and the equivalent terms are translations of: the identified terms; more general terms than the identified terms; and more specific terms than the identified terms.

5. A method as claimed in claim 4, characterised in that the multilingual resource (11) comprises a glosser (13).

6. A method as claimed in claim 5, characterised in that the glosser (13) is a limited non-deterministic glosser.

7. A method as claimed in claim 6, characterised in that the glosser (13) forms a plurality of translations of at least one of the identified terms and assigns to each translation a priority according to the likelihood of the translation being correct.

8. A method as claimed in claim 4, characterised in that the multilingual resource (11) comprises a bilingual dictionary (15).

9. A method as claimed in claim 4, characterised in that the multilingual resource (11) comprises a machine translation system (14).

10. A method as claimed in any one of the preceding claims, characterised in that the identifying step (24, 25) is performed by a part of speech tagger (16).

11. An apparatus for forming, for a plurality of documents (1, 23), an index comprising indexing features, characterised by comprising:

    means (24, 25) for identifying each of at least some of the terms present in the documents (1, 23);

means (28) for generating form each identified t rm at least one equivalent term which is different from but linguistically related to the identified term;

means (34) for forming for each of the id ntifi d terms a first indexing featur comprising the identified term and an identifier of the or each document in which the identified term occurs.

means (34) for forming for each of the equivalent terms a second indexing feature comprising the equivalent term and an identifier of the or each document in which the identified term to which the equivalent term is equivalent occurs; and

means for forming an index comprising the first and second indexing features.

12. An apparatus as claimed in claim 11, characterised in that the documents (1, 23) are natural language documents.

13. An apparatus as claimed in claim 11 or 12, characterised in that the generating means (28) comprises a thesaurus (12) and the equivalent terms are synonyms of the identified terms, more general terms than the identified terms and more specific terms than the identified terms.

14. An apparatus as claimed in claim 11 or 12, characterised in that the identifying means (24, 25) and the generating means (28) comprise a multilingual resource (11).

15. An apparatus as claimed in claim 14, characterised in that the multilingual resource (11) comprises a glosser (13).

16. An apparatus as claimed in claim 15, characterised in that the glosser (13) is a limited non-deterministic glosser.

17. An apparatus as claimed in claim 16, characterised in that the glosser (13) is arranged to form a plurality of translations of at least one of the identified terms and to assign to each translation a priority according to the likelihood of the translation being correct.

18. An apparatus as claimed in claim 14, characterised in that the multilingual resource (11) comprises a machine translation system (14).

19. An apparatus as claimed in claim 11 or 12, characterised in that the generating means (28) comprises a bilingual dictionary (15).

20. An apparatus as claimed in any one of claims 11, 12 and 19, characterised in that the identifying means (24, 25) comprises a part of speech tagger (16).

21. An apparatus as claimed in any one of claims 11 to 20, characterised by comprising a programmed data processor (2, 7).

22. A storage medium (7) characterised by containing a program for controlling a data processor (2) to perform a method as claimed in any one of claims 1 to 10.

23. An index characterised by being formed by a method as claimed in any one of claims 1 to 10 or by an apparatus as claimed in any one of claims 11 to 21.

24. A storage medium (8, 9, 10) characterised by containing an index as claimed in claim 23.

25. Use of an index as claimed in claim 23 to access the documents (1, 23).

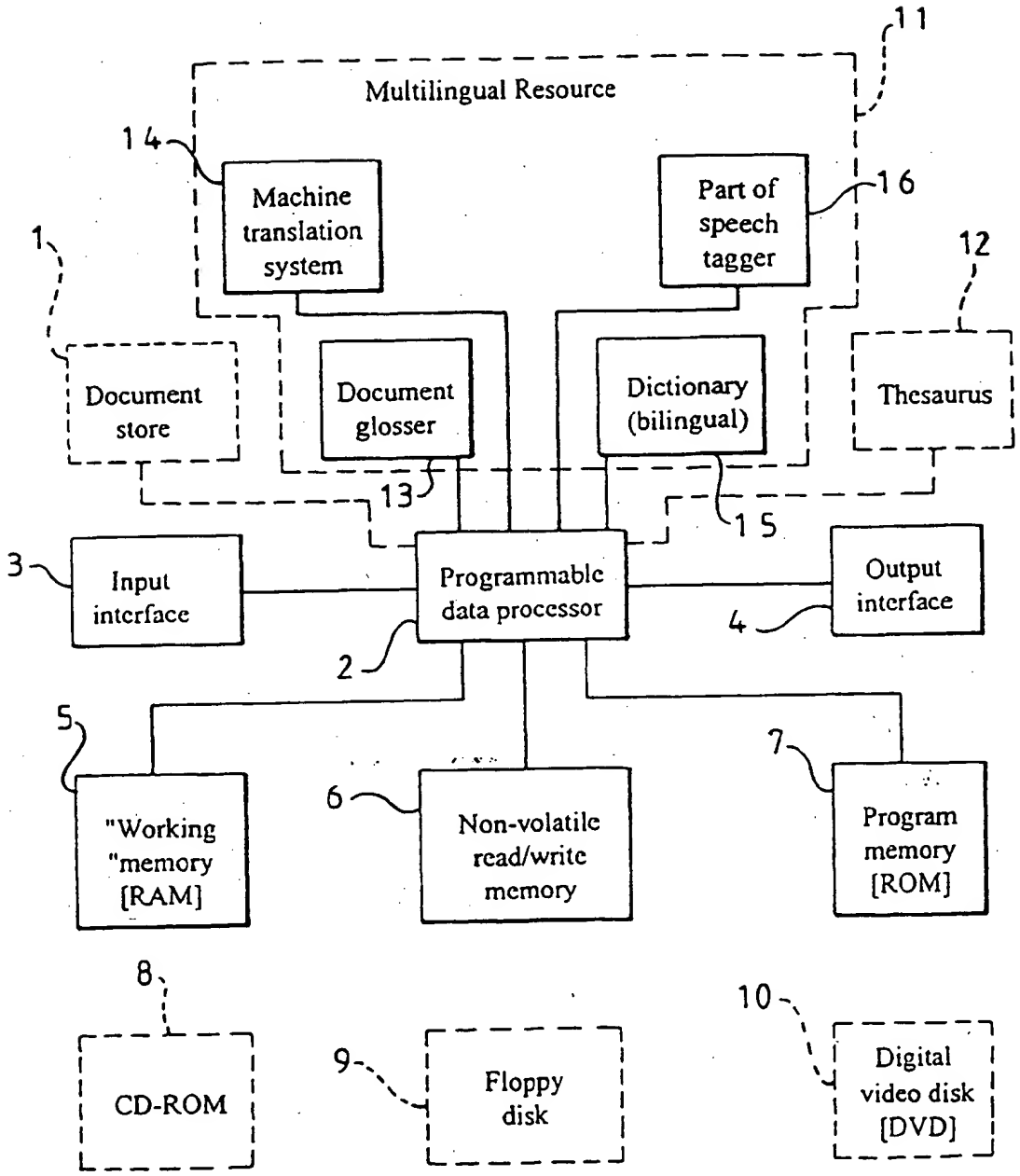

FIG 1

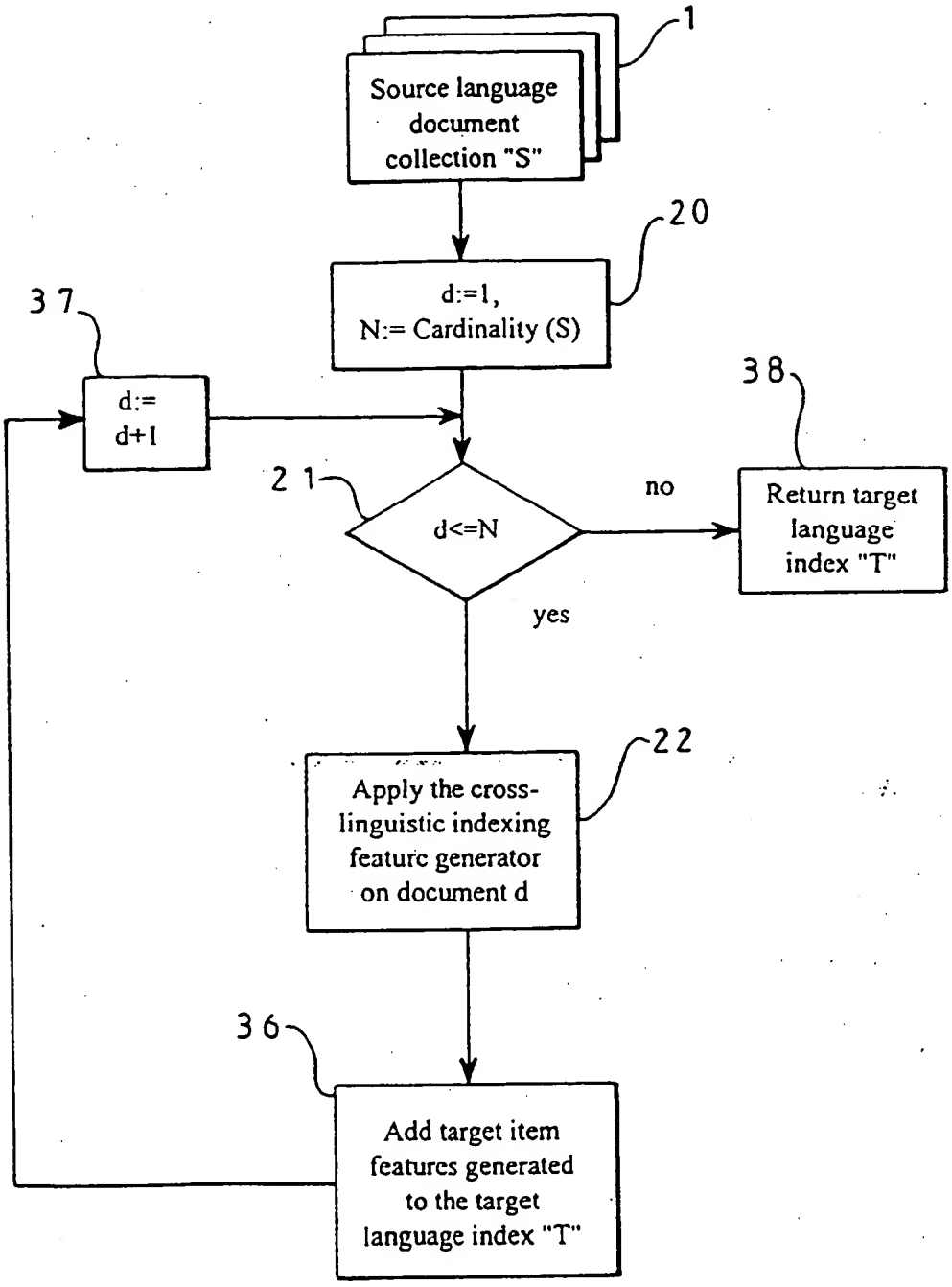

FIG 2

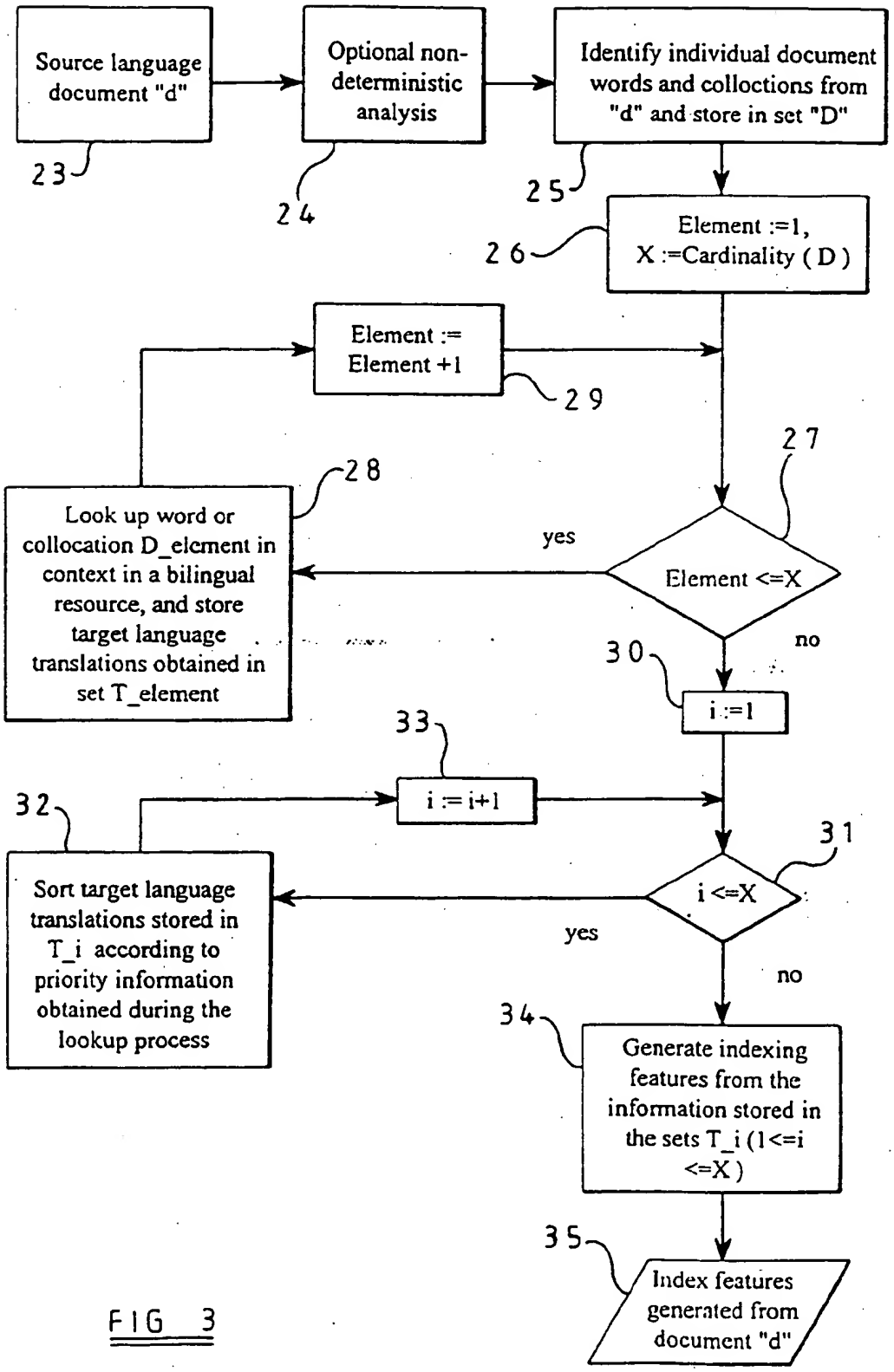

FIG 3

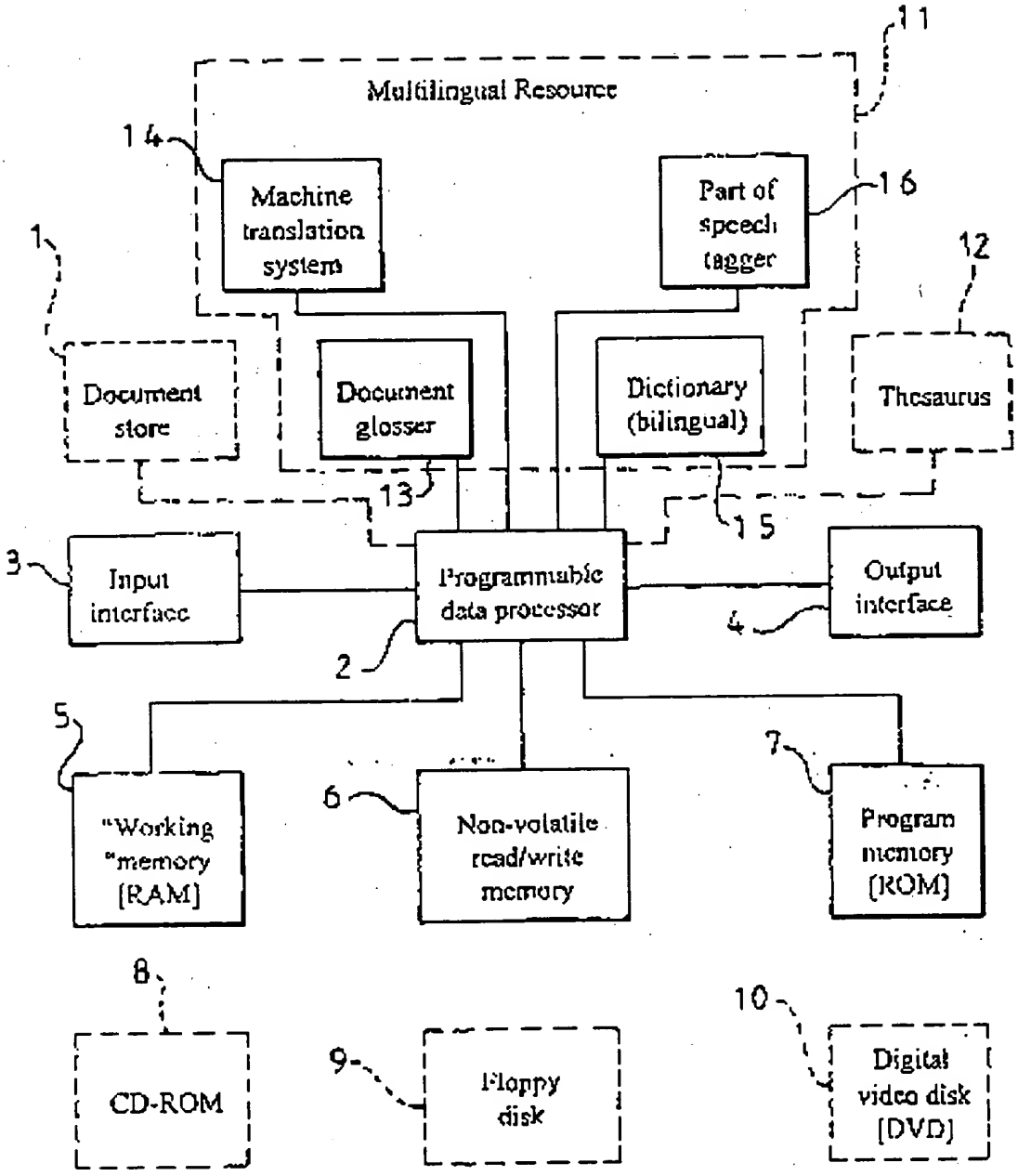

FIG 1

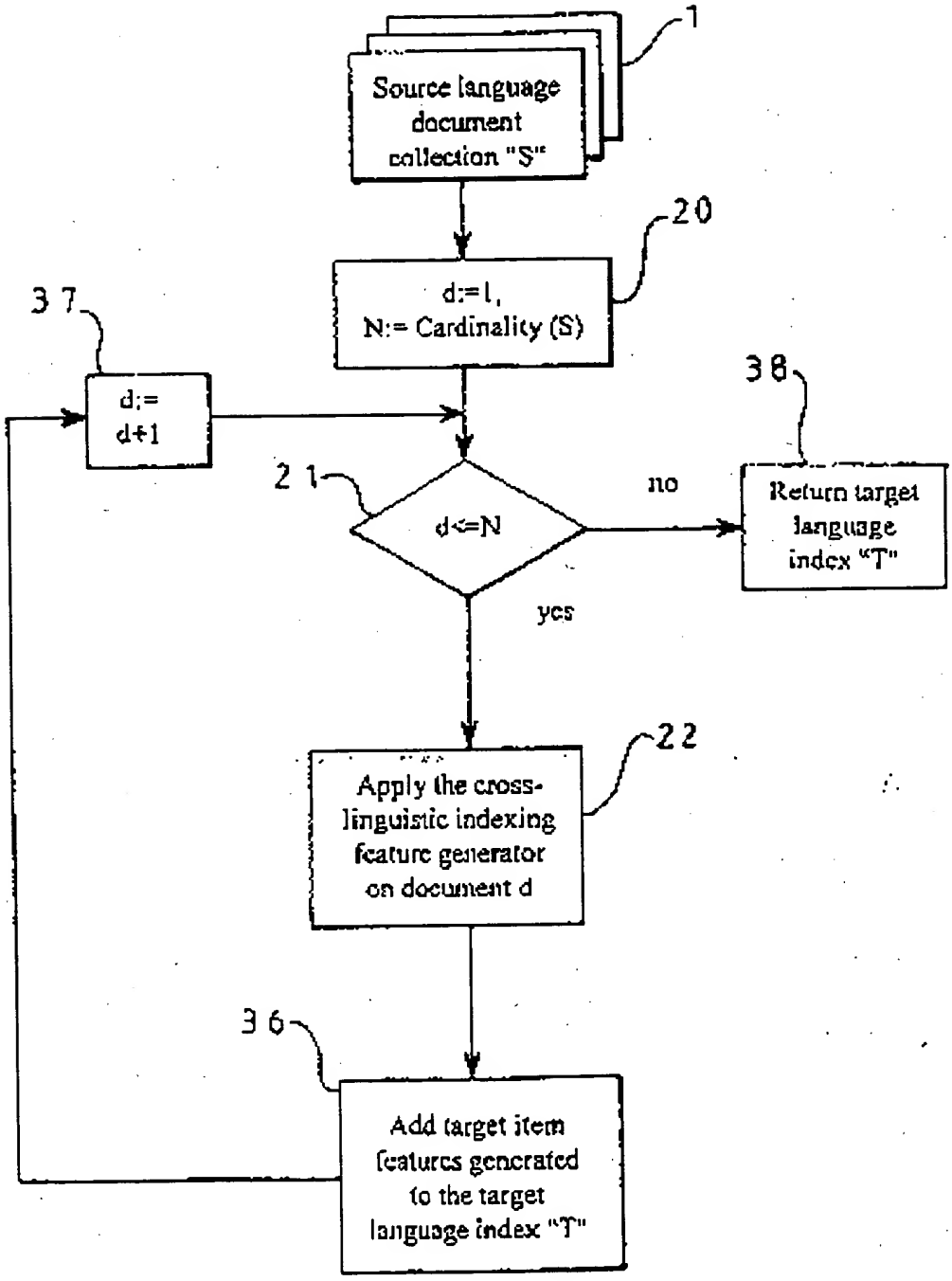

FIG 2

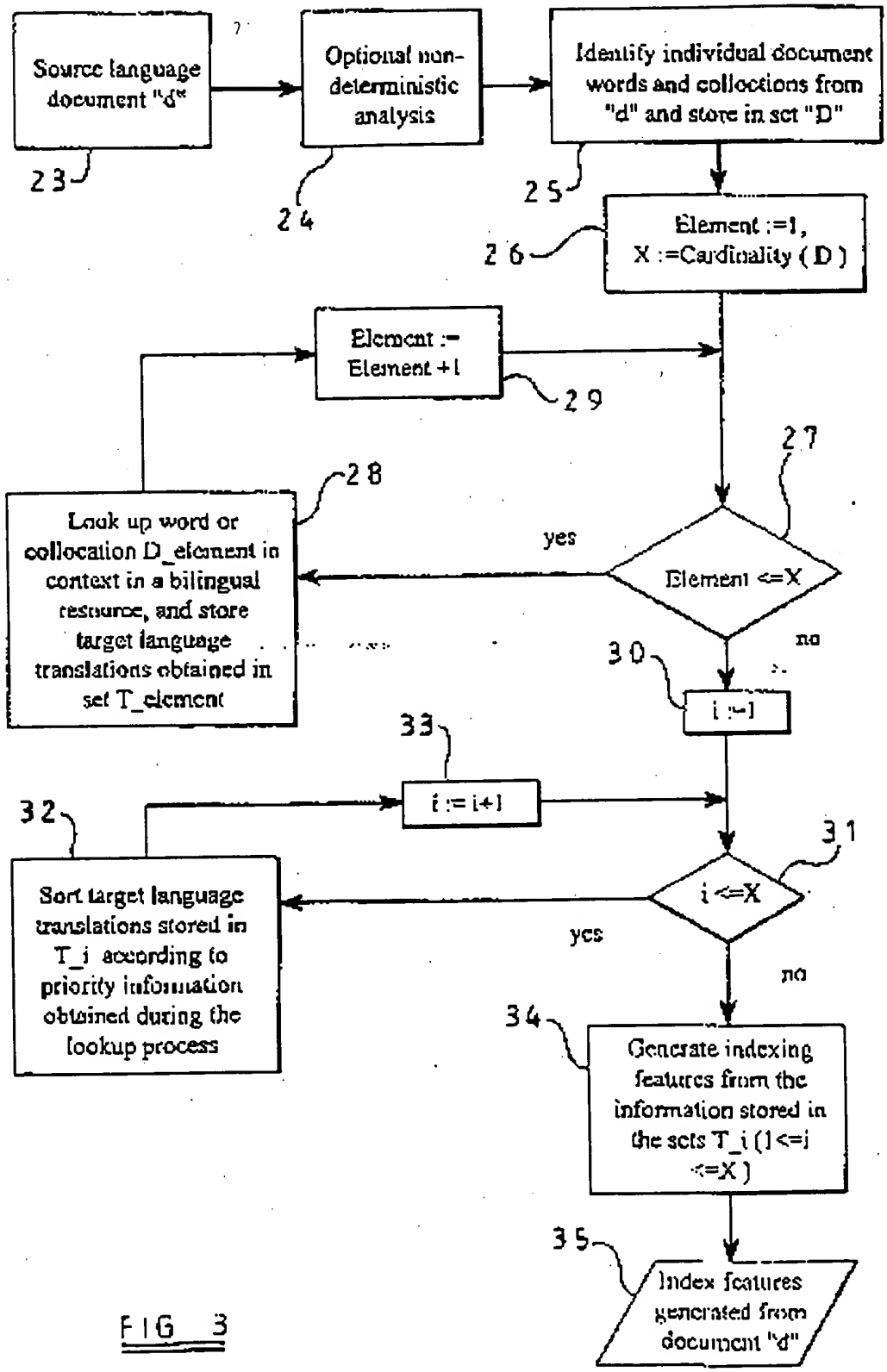

FIG 3